

YASH RAMESH SHELAR

linkedin.com/in/shelar-yash | github.com/YashShelar007 | yashshelar.com

SKILLS

Cloud & Infrastructure: AWS (ECS Fargate, Lambda, S3, ALB, DynamoDB, Step Functions, IAM Identity Center), GCP, Azure, Snowflake, Docker, Kubernetes, Pulumi, Terraform (IaC), CI/CD Pipelines, Linux, Microservices, System Design.

Programming & Scripting: Python, Go, TypeScript, JavaScript, Shell/Bash, SQL, C, C++, Java.

AI/ML & LLM Integration: LangChain, LangGraph, Model Context Protocol (MCP), OpenAI API, Anthropic API, RAG Architectures, Multi-agent Workflows, LLM Evaluation & LLM-as-Judge, pgvector, PyTorch, TensorFlow.

Observability & Deployment: CloudWatch, Langfuse, Sentry, GitHub Actions, Vercel, Container Orchestration, Secrets Management.

Frameworks & Databases: Angular, FastAPI, Flask, React, Next.js, Node.js, PostgreSQL, MongoDB, MySQL, Redis, Supabase.

WORK EXPERIENCE

Software Development Engineer | Walnutech PBC

Sep 2025 to present

Converted to full-time April 2026.

- Built **Vantion**, the company's product, from zero to **200+ users**; personally responsible for **~90% of implementation** across backend, ML, infrastructure, console and frontend.
- Built **LLM evaluation harnesses** in production (LLM-as-judge with calibrated scoring and regrounding, multi-turn A/B evaluation, and a reranker calibration harness), turning subjective output quality into reproducible signal.
- Ran **model-selection studies** that drove production decisions, including selecting Haiku over Sonnet on a measured quality/cost tradeoff rather than intuition.
- Authored an **organization-wide AI-native engineering toolkit**: 15 workflow skills, custom agents, and safety hooks distributed to **8 repositories** via an internal plugin marketplace; audited every repo on a 0 to 5 agent-readiness rubric and briefed leadership on the rollout.
- Designed and implemented **cloud-native infrastructure** across a multi-account AWS organization (Dev/Test/Prod) with IAM Identity Center, SCP guardrails, and environment-specific roles, **reducing access requests by 65%**.
- Replaced a 3GB MindsDB sidecar with a pgvector hybrid search engine** over **43,000+ scholarships**, cutting average query latency by **5x**, eliminating **120s bootstrap time**, and reclaiming **4GB memory per ECS task**.
- Stack: Python, FastAPI, LangChain/LangGraph, Langfuse, pgvector, MCP, OpenAI/Anthropic/Tavily, Docker/ECR, AWS (ECS Fargate, ALB, CloudWatch, IAM Identity Center, SSM/Secrets Manager), GCP, Terraform/Copilot, GitHub Actions (OIDC), Poetry/Pytest.

Software Developer | United Portraits

Jan 2026 to Mar 2026

- Built **SMS Manager**, a Twilio messaging platform with batch scheduling, opt-out tracking, and quiet-hours enforcement, raising the organization's **Twilio compliance score from 78% to 94%** within 4 weeks and **98.6% weekly compliance**.
- Stack: Angular 21, TypeScript, Fabric.js, Three.js, RxJS, Spring Boot 4, Java 21, Twilio, Keycloak, Docker Compose, Nginx, PostgreSQL, OpenAI API, Apache PDFBox, SCSS Design System.

Cloud & ML Researcher | VISA Research Lab

Jan 2024 to Aug 2025

- Engineered a serverless **infrastructure automation** pipeline (MCaaS) on AWS that scaled to **100+ executions/month**, reducing manual ML compression effort by **90%** and deployment failures by **42%**.
- Orchestrated distributed workflows using **AWS Step Functions**, achieving **14x model size reduction** and **3.4x faster inference** with less than **4% accuracy loss**.
- Stack: Python, Bash, PyTorch, AWS (ECS Fargate, S3, Lambda, Step Functions), Docker, Terraform, Git.

RELEVANT PROJECTS

pathwise: Trajectory Evaluation for AI Agents

Aug 2026

Open-source, Apache-2.0, archived.

- A framework for evaluating **how** an agent reached an answer rather than only the output, catching wrong reasoning, tool misuse, loops, and silently recovered errors that output-level evals pass.
- Treats agents as **stochastic**: assertions run across N repeated runs and report a pass-rate distribution rather than a single boolean, which is what makes prompt and model changes measurable.

Friday: Agentic Personal Assistant, Production PWA

Aug 2026 to present

- Built and deployed an **LLM assistant with tool use** (journal retrieval over Postgres full-text search, web search with citations, application tracking, email drafting) as a Next.js PWA with Web Push and scheduled jobs.
- Designed a **three-tier tool permission model** (read/write/act) enforced in one execution path, so any irreversible or outward-facing action requires explicit confirmation before it will run. Email is drafted, never sent.
- Instrumented **per-call cost accounting**: tokens recorded as measured fact, dollar cost as a versioned estimate, and **NULL rather than zero** for an unpriced model, so a pricing correction re-derives history instead of silently understating spend.
- Enforced operational rules **mechanically** rather than in documentation: pre-execution hooks that refuse a production deploy from the CLI and refuse any database write against the wrong project, each added after that failure mode occurred once.

EDUCATION

MS Computer Science, Arizona State University, Tempe AZ, Aug 2025. GPA 3.22. Thesis: Model Compression as a Service for Resource-Constrained Edge Devices.

BS Computer Science, Arizona State University, Tempe AZ, Dec 2022. GPA 3.92. Summa Cum Laude, New American Scholar, Dean's List. Relevant coursework: Cloud Computing, Distributed Database Systems, Machine Learning, Artificial Intelligence, Operating Systems.